融合图片相似度缓解新项目冷启动问题的研究 *

周　强，胡　燕

(武汉理工大学 计算机科学与技术学院，武汉 430070)

摘　要：针对推荐系统中因新项目的加入而造成的冷启动问题，在矩阵分解模型的基础上提出了融合项目图片相似度和类别属性的协同过滤推荐模型 USPTMF-CFIA。首先，采用基于用户偏好和时间权重的矩阵分解模型，对评分缺失项进行预测填充；然后，利用 VGG16 神经网络提取项目图片特征，并结合类别属性计算新项目与历史项目的相似度，得到近邻项目；最后，根据新项目与近邻项目之间的相似度预测用户对新项目的评分，将评分高的前 N 个项目推荐给对应用户；通过在 GroupLens 提供的数据集上的实验证明，该模型的推荐准确率比 MAP-BPR 模型高 0.006~0.015，比传统协同过滤模型高 0.02~0.028，比没融合图片相似度的 USPTMF-CFA 模型高 0.001~0.003，比 ACMF 模型高 0.001~0.002。

关键词：协同过滤；矩阵分解；图片特征；新项目冷启动；时间权重

中图分类号：TP391.1　　doi: 10.3969/j.issn.1001-3695.2018.01.0117

Research on solving new item cold-start problem by combining image similarity

Zhou Qiang, Hu Yan

(*College of Computer Science & Technology*, *Wuhan University of Technology*, *Wuhan* 430070, *China*)

Abstract: Aiming at the problem of cold start caused by the addition of new item in the recommendation system, This paper proposed a collaborative filtering recommendation model USPTMF-CFIA based on matrix factorization model, which combines the similarity of item image and category attributes . First, it used the matrix factorization model based on users' preference and time weight to predict and fill the missing item. Then, it used the VGG16 neural network to extract the features of the item images and combines category attributes to calculate the similarity between the new item and the historical items, then got the item's neighbors. Finally, the new item is predicted based on the similarity between the new item and the neighbors, and the first N items with high score are recommended to the correspond user. The experiment on the dataset provided by GroupLens proved that the proposed accuracy rate of this model. The recommended accuracy of this model is 0.006~0.015 higher than the MAP-BPR model , 0.02~0.028 higher than the traditional collaborative filtering model and 0.001 ~ 0.003 higher than that of the USPTMF-CFA model without image similarity0.001~0.002 higher than ACMF model.

Key words: collaborative filtering; matrix factorization; image features; new item cold-start; time weight

0　引言

互联网的发展迅猛，社交、视频、广告、音乐、外卖、电商等领域越来越普及，已经渗透到了人们生活的方方面面。如何给用户提供更高品质的服务，满足用户真正的需求，提升用户体验，是现在很多系统普遍关注的问题。推荐系统是现如今最流行的技术，根据用户行为日志分析用户偏好，自动为用户推荐感兴趣的项目[1]。目前推荐系统核心算法仍然是协同过滤，由于其简单有效，已被广泛的研究与应用。其原理是物以类聚人以群分，通过计算当前用户（项目）的 KNN(K-Nearest-Neighbor)最近邻，然后根据近邻用户（项目）的评分记录来预测用户对项目的评分，根据评分高低来决定是否将该项目推荐给用户，可以看出协同过滤对历史数据具有很强的依赖。对于新用户、新项目、新系统来说，由于没有任何历史数据，导致推荐效果不佳的问题，我们称之为冷启动问题。

因此，本文针对协同过滤算法中存在的问题之一——新项目冷启动问题展开了研究。本文的主要工作：

a)根据艾宾浩斯遗忘曲线[2]函数，提出了符合人们兴趣变化的时间权重函数。并将其融入到用户偏好的矩阵分解模型（Matrix Factorization Recommendation Algorithm Based On User's Preference And Time Weight，简称 USPTMF）中，使得该模型能预测出更贴近用户兴趣变化的缺失评分项，缓解了协同

收稿日期：2018-01-23；**修回日期：**2018-03-15　　**基金项目：**湖北省自然科学基金重点类项目（2017CFA012）

作者简介：周强（1992-），男，湖北荆门人，硕士研究生，主要研究方向为数据挖掘、机器学习、信息检索（1175192850@qq.com）；胡燕（1965-），女，湖北松滋人，教授，博士，主要研究方向为数据挖掘、信息检索.

过滤因数据稀疏性导致推荐效果不佳的问题，提高了推荐准确性。

b)首先分析可用的项目属性，然后选取项目特征属性，证明了图片属性对于项目区分的重要性，再对项目图片进行特征提取，计算图片特征相似度(image feature similarity，IFS)，结合项目类别属性相似度(attribute similarity，AS)得到 IAS 相似度，最终使得到的近邻项目更加精确。

c)通过 IAS 相似度计算出新项目的 k 近邻，并根据新项目与近邻项目集合之间的 IAS 相似度计算用户对新项目的预测评分，然后将得到的预测评分进行排序，取评分高的项目推荐给对应用户。将耗时的 USPTMF 模型训练放在线下，然后结合 KNN 算法进行在线推荐，相对于矩阵分解推荐模型提高了推荐实时性。

1　相关工作

推荐系统是决策者的有效工具，可以帮助他们根据自己的喜好挑选合适的项目，如今已经渗透到各个领域，但因新用户或新项目的加入，而导致的冷启动问题给各类推荐系统造成了很大困扰。为缓解推荐系统中的新项目冷启动问题，相关领域的专家和学者进行了大量的研究和努力。

为了缓解新项目冷启动问题，Hdioud 等人[3] 提出了基于项目混合特征选择方法，对项目内容属性进行聚类来缓解新项目冷启动问题。于洪等人[4] 利用项目属性、项目标签和用户时间权重，提出了一种新的计算项目相似度的方法，来缓解新项目冷启动问题。Aleksandrova 等人[5] 通过寻找种子用户，建立非种子用户与种子用户的联系，以种子用户对项目的评分为基础，来对新项目进行预测评分，其利用了线性代数中的基向量的思想，而且提出的模型具有一定的解释性。张栩晨等人[6] 在 Tri-training 框架的基础之上，提出了融合用户项目上下文信息的矩阵分解模型。Gantner 等人[7] 在基于贝叶斯个性化排序(Bayesian Personalized Ranking ,简称 BPR)框架的矩阵分解模型上,利用函数映射的关系，建立起类别属性到用户偏好之间的联系，来预测用户对新项目的偏好程度，以缓解新项目冷启动。任彩霞等人[8] 提出了利用信任网络和决策树来缓解新项目冷启动问题。Ocepek 等人[9] 将本地学习，属性选择和评分聚合运用在矩阵分解模型中提高新项目冷启动推荐精度，同时提高了推荐的可解释性。Wei 等人[10] 利用深度学习的方法提取的项目内容特征结合协同过滤预测新项目评分。Asmaa 等人[11] 通过提高项目类别属性相似度以及矩阵分解模型预测能力来缓解新项目冷启动问题。余永红等人[12] 通过改进项目类别属性的相似度——耦合相似度计算来缓解新项目冷启动。Wei 等人[13] 提出将协同过滤与深度学习相结合来解决冷启动问题。Barjasteh 等人[14] 和 Yuan 等人[15] 提出了一种基于矩阵分解的新算法，其同时利用用户和项目之间的相似性信息来缓解新项目冷启动。Saveski 等人[16] 利用项目的属性和历史用户偏好，提出了一种基于乘法更新规则的学习算法。Kula 等人[17] 提出了一种基于邻域的方法，构建用户和项目标签的关联矩阵，学习用户对标签的兴趣向量来预测评分。Geng 等人[18] 提出了一种深度学习模型，用来学习用户和图像的特征表示，缓解数据稀疏性问题。以上工作，基于项目的协同过滤算法能让推荐更加多元化，个性化，但是新项目没有任何评价信息，相似度计算无法使用。基于模型的协同过滤算法能解决传统协同过滤算法中数据稀疏的问题，但是解释性太差，个性化推荐质量不高。本文将结合两者的优势，基于 USPTMF 模型本文提出了融合项目图片相似度和类别属性的协同过滤推荐模型（USPTMF-collaborative filtering with project images and attributes，USPTMF-CFIA），有效地缓解了新项目冷启动问题。

2　问题定义及相关算法

2.1　数据定义

本文用 u 表示用户，i 表示项目，m 表示用户数，n 表示项目数，则用户集合表示为 $\mathbf{U}=\{u_1,u_2,u_3,\cdots,u_{m-1},u_m\}$，项目集合表示为 $\mathbf{I}=\{u_1,u_2,u_3,\cdots,u_{n-1},u_n\}$，用户 $u\in U$ 对项目 $i\in I$ 的评分表示为 R_{ui}，用户 $u\in U$ 对项目 $i\in I$ 的预测评分表示为 $\hat{R}_{ui}$，真实用户-项目评分矩阵表示为 $\mathbf{D}$，预测用户-项目评分矩阵表示为 $\mathbf{D}'$，所有项的平均评分表示为 μ 。

2.2　时间权重

随着时间的推移，用户的兴趣也会随之发生变化，这些变化体现在用户不断增加的新行为中，评论时间越近更能反映用户的兴趣，对评分效应越大，时间权重因子越大[19]，而且不同的年龄段的用户，其兴趣爱好也不一样，假设用户年龄越大记忆越弱，兴趣衰减越快。根据尹三文[20] 对数据集 Netflix 中评分时间跨度较大的三个用户进行了分析，这三个用户的平均评分很符合德国心理学家艾宾浩斯提出的人类大脑对新事物的遗忘规律[2]，据此本文提出了如下时间权重函数形式，记为 $t(u,i)$：

$$t(u,i)=e^{\frac{-a(t_{\max}-t_{ui})}{a_{\max}(t_{\max}-t_{\min})}} \tag{1}$$

其中：t_{ui} 表示用户 u 对项目 i 的评论时间，$t_{\max}$ 表示用户 u 最近评论时间，$t_{\min}$ 表示用户 u 最近评论时间，a 表示用户 u 的年龄，$a_{\max}$ 表示所有评论用户最大年龄。

从式(1)可以看出，当一个评分是最先评论的时候，$w(u,i)=e^{\frac{-a}{A_{\max}}}$，根据 Movielens 数据集[21] 中提供的用户信息，

用户的最大年龄为 73 岁和最小年龄为 7 岁，可以计算出时间权重 $w(u,i)\in[0.3678,0.9085]$，利用(1)式能很好的计算出不同年龄段的用户，时间权重对评分的影响，年纪越大记忆力可能越弱，符合大部分人类的情况。当一个评分是最近评论的时候，$w(u,i)=1$，评分时间最近最能反应当前用户的兴趣走向，所以这时要赋予高权重，评分时间越久权重越小。

2.3 融合用户偏好和时间权重的矩阵分解模型

矩阵分解[22-23] 是将大的评分矩阵分解为小的矩阵，然后通过随机梯度下降法不断的迭代使子矩阵的乘积不断地逼近真实矩阵。我们希望预测评分 $\hat{R}_{ui}$ 和真实评分 R_{ui} 之间的误差越小越好。通过计算 $e_{ui}=R_{ui}-\hat{R}_{ui}$ 得到预测值和真实值的误差。让预测评分矩阵和真实评分矩阵的尽可能的接近，即让误差 e_{ui}^2 尽可能的小。由于每个用户携带有与事物无关的属性，有的用户宽容，给的评分相对较高，有的用户则比较苛责，所给评分相对较低，因此模型中加入了用户偏好偏置值，项目偏好偏置。此时预测评分 $\hat{R}_{ui}$ 表示为

$$\hat{R}_{ui}=\mu+b_u+b_i+\boldsymbol{p}_u^T\boldsymbol{q}_i \tag{2}$$

将 2.2 节中提到的时间权重函数加入到目标函数中，得:

$$E=\sum_{u,i\in \boldsymbol{D}}t(u,i)\left(R_{ui}-\hat{R}_{ui}\right)^2+\lambda\left(\|p_u\|^2+\|q_i\|^2+\|b_u\|^2+\|b_i\|^2\right) \tag{3}$$

其中：$\boldsymbol{p}_u$，$\boldsymbol{q}_i$ 分别表示用户和项目特征向量，b_u，b_i 分别表示用户和项目的偏好偏执值，λ 为正则化系数，防止过拟合。

2.4 项目相似度计算

随着电影库中数据的积累，同种类型的电影越来越多，我们如何根据电影类型来区分电影？很多研究都是改进类别属性相似度计算来更加精确地找到近邻项目，但是对同种类型的电影进行类别属性相似度计算，最终得到的相似度都为 1，取前 N 项时就会导致很多电影得不到推荐。所以本文将项目的“外在”特征——图片与目的“内在”特征——类别属性相结合计算项目的综合特征相似度，记为 IAS。

2.4.1 项目类别属性相似度

在实际推荐应用场景中，项目通常由类别属性值描述。以电影为例，电影有喜剧、动作、冒险、恐怖等类别属性。因此很多相关专家学者利用类别属性，对相似度进行改进来提高模型的预测准确度。由于项目—属性矩阵的庞大，采用大量的乘法运算将会耗费大量的时间，余永红等人[12] 提出的耦合相似度计算虽然能有效提高预测准确度，但是计算量相当大。在保证可行性的条件下，提高时间效率非常重要。一般情况下，为了方便计算项目之间的相似度，不仅要构建用户-项目评分矩阵，还要提取项目的类别属性矩阵。简单的示例如表 1、2 所示。

表 1 用户-项目评分矩阵

	I1	I2	I3	I4	5
U1	1	3	4	1	5
2	2	2	5	1	4
U3	1	3	5	2	3
U4	2	1	4	1	5

表 2 项目-类别属性矩阵

	A1	A2	A3	A4	A5
I1	1	1	1	0	1
I2	1	1	0	0	0
I3	0	1	0	0	1
I4	0	1	1	0	1
I5	1	0	0	1	1

对于类别属性的数据，通常利用 0 和 1 作余弦相似度计算。如表 1 所示，I1 的类别属性集合为{1，1，1，0，1}，I3 的类别属性集合为{0，1，0，0，1}，通过类别集合计算 I1 和 I3 两者之间的相似度为 0.707，但分析表 1 可以知道，I3 的平均分为 4.5， I1 的平均分为 1.5，说明影片 3 远比影片 1 好看，更加受到用户的喜爱，但是得到两者之间的相似度有点差强人意；再看一个例子，I4 的类别属性为{0，1，1，0，0}，通过类别集合计算 I3 和 I4 的相似度为 0.816，反观影片 4 的评分远远低于影片 3 的评分，直接根据类别属性计算两者之间的相似度这种方法粒度太粗。分析表 2，出现的频次有多有少，属性 A2 和属性 A5 出现的次数较为频繁，属性 A4 出现的频次较少，那么属性出现的频次是否会影响到项目相似性的计算呢？对于电影来说，影片可以分为喜剧和悲剧，喜剧分为讽刺喜剧、欢乐喜剧、幽默喜剧、无厘头喜剧等小类，因此大类属性-喜剧这个属性出现的频次相比于其小类出现的频次要高，那么在计算相似度时赋予每种属性的权重应该有所不同。人们常说，“物以稀为贵”，所以出现频次少的属性应该赋予更高的权重。为了提高计算效率，很容易联想到将属性出现的总频次的倒数作为权重值，于是得到如表 3 所示的矩阵。

表 3 引入权重的项目-类别属性矩阵

	A1	A2	A3	A4	A5
I1	0.33	0.25	0.5	0	0.25
I2	0.33	0.25	0	0	0
I3	0	0.25	0	0	0.25
I4	0	0.25	0.5	0	0.25
I5	0.33	0	0	1	0.25

这时，I1 的类别属性集合为{0.33，0.25，0.5，0，0.25}，I3 的类别属性集合为{0，0.25，0，0，0.25}，I4 的类别属性集合为{0，0.25，0.5，0，0.25}，得到影片 1 和影片 3 的相似度为 0.595，相比于 0.707 更容易让人接受；影片 3 和影片 4 的相

似度为 0.246 相比于 0.816 能够反映两部电影的差异，结果更加让人信服。

Movielens 数据集[21] 中项目的类别属性是由 0-1 组成的向量，比如 ml-100k 数据集中，电影《Toy Story》的类别属性向量为：$\boldsymbol{i}=\{0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0\}$，共有 19 位，计算 i_i, i_j 两个项目的相似度的公式如下：

$$sim_a(i_i,i_j)=cos(i_i,i_j)=\frac{\sum_{x=1}^{k}\frac{a_{ix}}{g(a_{ix})}*\frac{a_{jx}}{g(a_{jx})}}{\sqrt{\sum_{x=1}^{k}\frac{a_{ix}^2}{g(a_{ix})^2}*\sum_{x=1}^{k}\frac{a_{jx}^2}{g(a_{jx})^2}}} \tag{4}$$

其中：a_{ix} 表示项目 i_i 属性向量的第 x 位属性，$g(a_{ix})$ 表示属性 a_{ix} 出现的频次，k 表示属性向量 $\boldsymbol{i}$ 的长度。

2.4.2 项目图片相似度

现如今各大电商网站、新闻网站中各种各样精美的图片充斥着人们的眼球，打开页面，最先映入眼帘的就是图片，如何展现良好的第一印象给用户呢？那肯定是图片了。一张图片的好坏直接影响着商品(新闻)的点击率，项目的图片起着敲门砖的作用。从本质上来说，图片和文字一样，都是信息的载体。从项目图片中，在不知道项目属性信息情况下，可以大致预测这个项目的功能，电影就是娱乐和视觉艺术相结合绝佳的例子。人们通过海报可以知道电影的名称，预测电影的氛围场景，甚至可以通过电影海报发现电影类型。如图 1 所示，《Toy Story》海报该海报中的人物比较卡通，可以预测该电影的是动画类型的，可能受到小朋友喜欢，海报的颜色较为轻快明亮，可以预测出该电影可能是喜剧类型。



图 1 Toy Story 海报

还就电影《Toy Story》而言，该影片讲述了主角牛仔警长胡迪和太空骑警巴斯光年的故事，通过海报特征找到最相似的前 5 个项目(相似度从左向右依次降低)，如图 2 所示，这些海报中都有两个主要人物，同为动画类型。在真实数据集中，电影《Toy Story》的类型为动画、儿童、喜剧；电影《Chicken Run》的类型为动画、儿童、喜剧，与目标项目类型也相同；电影《A Grand Day Out》的类型为动画、喜剧，电影《Cats Don't Dance》的类型为动画、儿童、音乐，电影《Doug's 1st Movie》的类型为动画、儿童，电影《Jerry & Tom》的类型为戏剧。后面四部电影的类型与《Toy Story》不完全相同，这样可以提高推荐的新颖度，这是协同过滤所做不到的，因为协同过滤方法需要通过数据寻找属性类似的项目，导致了具有新属性的项目无法推荐。根据图片相似，我们可以得到较好的推荐效果。

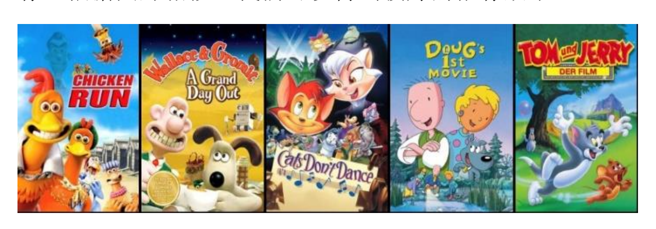


图 2 与《Toy Story》相似的项目

在实际推荐应用场景中，每个项目都会携带有图片元素，比如电影海报、商品展示图等。在 2.4.1 节中，我们把项目类别属性可以表示为一个一维的向量，那么图片能不能也表示成一个一维向量呢？如果能表示成一维向量我们就可以采用余弦相似度计算公式计算两张图片之间的相似程度。基于这种思想本文采用 VGG16 卷积神经网络对图片集进行特征提取，每张图片则表示成了一个长度为 25088 的一维向量 $\boldsymbol{p}$，其中元素为浮点数，计算两张图片的相似度公式如下：

$$sim_i(i_i,i_j)=cos(i_i,i_j)=\frac{\sum_{x=1}^{k}p_{ix}*p_{jx}}{\sqrt{\sum_{x=1}^{k}p_{ix}^2*\sum_{x=1}^{k}p_{jx}^2}} \tag{5}$$

其中：p_{ix} 表示项目 i_i 图片向量 $\boldsymbol{p}$ 的第 x 位数值，k 表示图片向量 $\boldsymbol{p}$ 的长度。

2.4.3 融合项目类别属性和图片特征的相似度计算

经过 2.4.1 节和 2.4.2 节的相似度计算，得到最终的融合项目类别属性和图片特征的相似度 IAS：计算公式如下：

$$sim_{ia}(i_i,i_j)=x*sim_i(i_i,i_j)+(1-x)*sim_a(i_i,i_j) \tag{6}$$

2.5 产生推荐

通过 2.3 节提出的 USPTMF 模型对用户评分缺失项进行预测填充，得到填充后的评分矩阵 D''，然后根据 IAS 相似度排序取前 N 项，根据新项目与近邻项目的相似度计算用户对新项目的预测评分，将评分高的前几个项目推荐给用户，预测用户评分的公式如下：

$$r_{ui}'=\frac{\sum_{n=1}^{N}sim_{ia}(i_i,i_j)*D_{uj}''}{\sum_{n=1}^{N}\left|sim_{ia}(i_i,i_j)\right|} \tag{7}$$

其中：D_{uj}'' 表示用户 u 对项目 j 的评分。

3 实验结果及分析

3.1 实验数据集

本文采用的是由美国 GroupLens 项目组提供的公开数据集

[21]，其中 ml-100k 数据集是由 943 个用户信息和 1682 部电影组成的。其中 u.data 文件包含 100000 条评分信息，u.user 文件包含 943 个用户信息，u.item 文件包含 1682 部电影信息。ml-1m 数据集是由 6040 个用户信息和 3706 部电影组成。其中 ratings.dat 文件包含 1000209 条评分记录，movies.dat 文件包含 3883 部电影信息，users.dat 文件包含 6040 个用户信息，两种数据集中用户评分都是 1 到 5 之间的整数。项目图片集是根据电影名查找电影的 imdb，利用 Movielens 提供的 api 下载。实验采用 5-fold 交叉验证，训练集所占比例为 80%，测试集所占比例为 20%。

3.2　实验预处理

3.2.1 构建测试集偏好矩阵

实验中，假设大于用户平均评分的整数部分的项，表示用户喜欢，记为 1，否则表示用户不喜欢，记为 0，这样就可以得到用户偏好矩阵。

3.2.2 冷启动项目的选取

将测试集中的评分数据全置为 0。

3.3　评价标准

本文采用 TopN 推荐准确度作为评价指标，验证本文改进算法的有效性。通过公式(7)来预测评分，提取前 N 个项目，根据推荐项目列表中某个被推荐的项目是否出现在了目标用户的测试集（3.2.1 节中提到的偏好集）中，判断是否生成了一个正确的推荐[24]，计算公式如下：

$$prec@n=\frac{1}{|\mathrm{U}_t|}\sum_{u\in U_t}\frac{|D_u^{''}(N)\cap T_{pu}|}{N} \tag{8}$$

其中：U_t 表示训练集中的用户集合，T_{pu} 为测试集用户偏好矩阵。

3.4　参数设置

本文提出的 USPTMF-CFIA 模型中包含多个参数，包括迭代次数 $iter$，学习率 $alpha$，正则化参数 λ，特征向量的维度 k。通过 5-folder 交叉验证，最终确认在 ml-100k 数据集上，$iter=30$，$alpha=0.001$，$k=40$，λ=0.01，在 ml-1m 数据集上，$iter=40$，$alpha=0.00001$，λ=0.01，$k=50$。

3.5　实验对比

为了验证本文提出算法的有效性，将与以下算法作对比：

a)USPMF：基于用户偏好的矩阵分解模型。

b)USPTMF：基于用户偏好和时间权重的矩阵分解模型。

c)Gantner 等人[7] 提出的 MAP-BPR 模型。

d)CBF-KNN：基于项目类别属性的协同过滤模型。

e)USPTMF-CFA：结合项目类别属性的 USPTMF 模型。

f)USPTMF-CFIA：融合项目图片相似度和类别属性的 USPTMF 模型。

g)ACMF：余永红等人[12] 提出的基于属性耦合的矩阵分解模型。

3.5.1 各个模型准确度对比

数据集 ml-100k 相比于数据集 ml-1m 小很多，所以本实验在数据集 ml-100k 中取所有用户作为评估对象，在 ml-1m 数据集中随机取 500 个用户作为评估对象，如表 4 为各个模型在数据集 ml-100k 中推荐 5 个项目的准确度（prec@5）对比，如表 5 为各个模型在数据集 ml-1m 中推荐 5 个项目的准确度(prec@5）对比。

表 4 ml-100k 数据集中各个模型准确度对比

Fold	USPMF	USPTMF	MAP-BPR	CBF-KNN	USPTMF-CFA	USPTMF-CFIA	ACMF
1	0.100106	0.106713	0.258536	0.137858	0.279745	0.288441	0.281212
2	0.037327	0.04369	0.064051	0.100318	0.115164	0.12492	0.116428
3	0.020361	0.029088	0.044963	0.058112	0.0386	0.043478	0.041585
4	0.007423	0.009544	0.025027	0.030328	0.022057	0.012937	0.022957
5	0.001697	0.00206	0.003393	0.003393	0.001485	0.003393	0.002475
Average	0.033383	0.038219	0.079194	0.066002	0.09141	**0.094634**	0.092931

表 5 ml-1m 数据集中各个模型准确度对比

Fold	USPMF	USPTMF	MAP-BPR	CBF-KNN	USPTMF-CFA	USPTMF-CFIA	ACMF
1	0.0312	0.0388	0.074	0.0656	0.11	0.1312	0.1242
2	0.0416	0.0436	0.1052	0.0696	0.1308	0.1312	0.1302
3	0.032	0.0412	0.0936	0.062	0.0756	0.09	0.0856
4	0.026	0.0354	0.1032	0.0708	0.0814	0.0576	0.0658
5	0.028	0.0304	0.0588	0.0564	0.0584	0.0528	0.0502
Average	0.03176	0.03788	0.08696	0.06488	0.09124	**0.09256**	0.0912

实验证明，前 5 项推荐实验中，模型 USPTMF-CFIA 预测正确率是最高的，模型 USPTMF 次之，说明加入图片特征，能

更加精确的找到近邻项目，有助于提高推荐精度。此外USPTMF相比于USPMF准确度提高了0.005~0.006，说明加入时间权重也能让矩阵分解模型预测更加准确；对于CBF-KNN模型，由于它采用协同过滤的方式，数据稀疏度会影响其推荐准确度；MAP-BPR模型比USPMF，USPTMF推荐正确率高，因为其结合了项目类别属性，而且MAP-BPR推荐实时性较高，但是该模型没有混入基于近邻的协同过滤思想，因此对于项目的其它不能编码的信息——图片就无法融入该模型，扩展性不高，所以没有本文提出的USPTMF-CFIA推荐准确度高；ACMF推荐准确度仅次于USPTMF-CFIA模型，但是耦合相似度的求解时间复杂度相当高。

3.5.2 邻居k对准确度的影响

如图3所示，模型USPTMF-CFA，USPTMF-CFIA在数据集ml-100k中推荐5个项目的准确度（prec@5）与邻居K的关系，如图4所示，模型USPTMF-CFA，USPTMF-CFIA在数据集ml-1m中推荐5个项目的准确度（prec@5）与邻居K的关系。

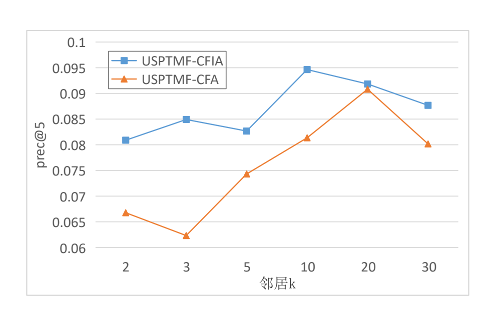


图3 ml-100k数据集中模型USPTMF-CFA，USPTMF-CFIA的prec@5与邻居K的关系

实验证明邻居数K取10的时候，模型USPTMF-CFIA在ml-100k数据集上prec@5最高，邻居数K取20的时候，模型USPTMF-CFA在ml-100k数据集上prec@5最高，但是低于模型USPTMF-CFIA。

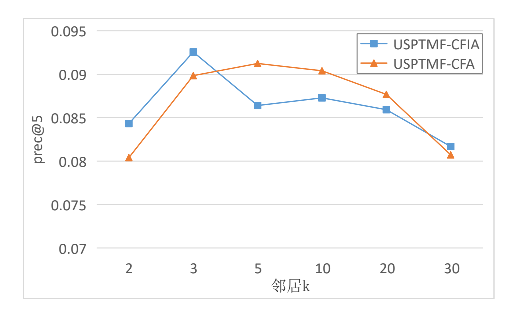


图4 ml-1m数据集中模型USPTMF-CFA，USPTMF-CFIA的prec@5与邻居K的关系

实验证明邻居数K取3的时候，模型USPTMF-CFIA在ml-1m数据集上prec@5最高，邻居数K取5的时候，模型USPTMF-CFA在ml-100k数据集上prec@5最高，但是低于模型USPTMF-CFIA。

3.5.3 式(6)中权重因子x对准确度的影响

如图5所示，模型USPTMF-CFIA在数据集ml-100k上，prec@5与公式(6)中权重因子x的关系，如图6所示，模型USPTMF-CFIA在数据集ml-100k上，prec@5与公式(6)中权重因子x的关系。

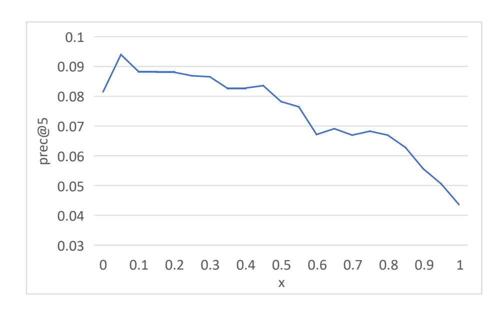


图5 模型USPTMF-CFIA在数据集ml-100k上，prec@5与公式(6)中权重因子x的关系

实验证明当 $x = 0.05$ 时，在数据集ml-100k上，prec@5最大，说明图片特征权重要低于属性特征，让项目属性占主导地位，同时也说明图片特征作为辅助信息目的就是为了区别属性相同的项目。

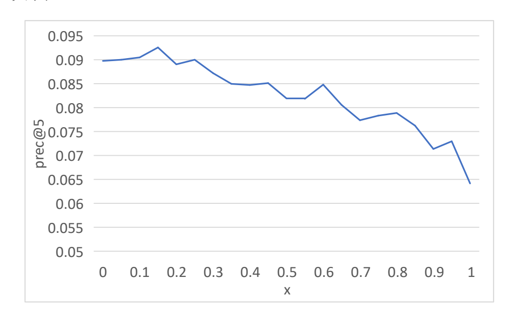


图6 模型USPTMF-CFIA在数据集ml-1m上，prec@5与公式(6)中权重因子x的关系

实验证明当 $x = 0.15$ 时，在数据集ml-1m上，prec@5最大。数据集ml-1m相比于数据集ml-100k，项目（电影）多了接近一倍，所以同类型的项目会更多，这时图片特征的作用将会增强，所以图片特征权重取值更大一些。

4 结束语

本文充分挖掘项目的特征信息找出近邻项目，同时将用户偏好和时间权重（用户因子）融入矩阵分解，将协同过滤与矩阵分解模型结合，融合图片相似度有利于提高冷启动项目推荐准确度。以往的很多研究都在改进项目相似度计算方式，比如余永红等人[12]提到的耦合相似度计算，对于项目数庞大的推荐系统，并不是最好的方案，虽然能很好计算项目之间的相似度，但是时间复杂度相当高，并且对于属性相同的项目推荐效果也不佳，这时结合项目的图片特征，能够更好的找出近邻项目，通过在数据集ml-100k和ml-1m上的实验证明，本文提出的模型能够有效缓解新项目冷启动问题，针对含有文字少、图片多的系统，能够产生很好的推荐效果。在接下来的工作是研究项目属性计算效率，同时将利用上下文关系，找出用户与项目之间的关系，来进一步提高推荐效果。

参考文献：

[1] Liu Chang, Wang Yulong. Analysis on the cold-start problem in recommendation system [J]. Telecommunications Network Technology, 2017 (1): 65-68.

[2] 于洪，李转运. 基于遗忘曲线的协同过滤推荐算法 [J]. 2010, 46 (5): 520-527. (Yu Hong, Li Zhuanyun. A collaborative filtering recommendation algorithm based on forgetting curve [J]. Journal of Nanjing University, 2010, 46 (5): 520-527.)

[3] Hdioud F, Frikh B, Benghabrit A, *et al.* Collaborative filtering with hybrid clustering integrated method to address new-item cold-start problem [M]// Intelligent Distributed Computing IX. [S. l] : Springer International Publishing, 2016: 285-296.

[4] 于洪，李俊华. 一种解决新项目冷启动问题的推荐算法 [J]. 软件学报, 2015, 26 (6): 1395-1408. (Yu Hong, Li Junhua. Algorithm to solve the cold-start problem in new item recommendations [J]. Journal of Software, 2015, 26 (6): 1395-1408.)

[5] Aleksandrova M, Brun A, Boyer A, *et al.* Identifying representative users in matrix factorization-based recommender systems: application to solving the content-less new item cold-start problem [J]. Journal of Intelligent Information Systems, 2017, 48 (2): 1-33.

[6] 张栩晨. 利用 Tri-training 算法解决推荐系统冷启动问题 [J]. 计算机科学, 2016, 43 (12): 108-114. (Zhang Xuchen. Utilizing tri-training algorithm to solve cold start problem in recommender system [J]. Computer Science, 2016, 43 (12): 108-114.)

[7] Gantner Z, Drumond L, Freudenthaler C, *et al.* Learning attribute-to-feature mappings for cold-start recommendations [C]// Proc of IEEE International Conference on Data Mining. IEEE Computer Society. 2010: 176-185.

[8] Ren Caixia. Improved algorithm of alleviating Item cold starting [J]. Computer Engineering &Software, 2016, 37 (8): 11-15

[9] Ocepek U, Rugelj J, Bosnić Z. Improving matrix factorization recommendations for examples in cold start [J]. Expert Systems with Applications, 2015, 42 (19): 6784-6794.

[10] Wei Jian, He Jianhua, Chen Kai, *et al.* Collaborative filtering and deep learning based recommendation system for cold start items [J]. Expert Systems with Applications, 2016, 69 (3): 29-39.

[11] Elbadrawy A, Karypis G. User-specific feature-based similarity models for top-n recommendation of new items [J]. ACM Trans on Intelligent Systems & Technology, 2015, 6 (3): 1-20.

[12] 余永红. 融合多源信息的推荐算法研究 [D]. 南京：南京大学, 2017. (Yu Yonghong. Research on recommendation algorithms via incorporating side information [D]. Nanjing: Nanjing University, 2017) .

[13] Wei Jian, He Jianhua, Chen Kai, *et al.* Collaborative filtering and de-ep learning based hybrid recommendation for cold start problem [C]// Proc of the 14th International Conference on Pervasive Intelligence and Computing, the 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress. 2016: 874-877

[14] Barjasteh I, Forsati R, Masrour F, *et al.* Cold-start item and user recommendation with decoupled completion and transduction [C]// Proc of ACM Conference on Recommender Systems. New York: ACM Press, 2015: 91-98.

[15] Yuan Jianbo, Shalaby W, Korayem M, *et al.* Solving cold-start problem in large-scale recommendation engines: a deep learning approach [C]// Proc of IEEE International Conference on Big Data. 2017: 1901-1910.

[16] Saveski M, Mantrach A. Item cold-start recommendations [C]// Proc of the 8th ACM Conference on Recommender Systems. NewYork: ACM Press, 2014: 89-96.

[17] Ji Ke, Shen Hong. Addressing cold-start: scalable recommendation with tags and keywords [J]. Knowledge-Based Systems, 2015, 83 (1): 42-50.

[18] Geng Xue, Zhang Hanwang, Bian Jingwen, *et al.* Learning Image and User Features for Recommendation in Social Networks [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 4274-4282.

[19] Liu Yun, Wang Ying, Qi Guotao. Recommendation Algorithm Based on Time Context [J]. Computer Technology & Development. 2017, 27 (7): 79-82.

[20] 尹三文. 一个改进的推荐算法研究与应用 [D]. 武汉：华中科技大学, 2015. (Yi Sanwen. Study and Application on an Improved Recommendation Algorithm [D]. Wuhan: Huazhong University of Science and Technology)

[21] GroupLens [EB/OL]. http://grouplens. org/datasets/movielens/.

[22] Wen Hailong, Ding Guiguang, Liu Cong, *et al.* Matrix factorization meets cosine similarity: addressing sparsity problem in collaborative filtering recommender system [C]// Proc of Asia-Pacific Web Conference. Springer International Publishing, 2014: 306-317.

[23] Liu Nathan N, Cao Bin, Zhao Min, *et al.* Adapting neighborhood and matrix factorization models for context aware recommendation [M]. New York: ACM Press, 2010: 7-13.

[24] 刘江冬，梁刚，杨进. 基于时效性的冷启动解决算法 [J]. 现代计算机, 2016 (5): 3-6. (Liu Jiangdong, Liang Gang, Yang Jin. Timeliness-Based Algorithm for Cold Start [J]. Modern Computer, 2016 (5): 3-6.)